UCLA Statistics Series
Report No. 197

Some New Test Statistics for Mean and Covariance Structure Analysis with High Dimensional Data

Ke-Hai Yuan and Peter M. Bentler
University of California, Los Angeles

November 7, 1995

This work was supported by National Institute on Drug Abuse Grants DA01070 and DA00017. The ideas developed in this paper were improved through discussions with Wai Chan and Robert I. Jennrich.

Abstract

Covariance structure analysis is often used for inference and for dimension reduction with high dimensional data. When data is not normally distributed, the asymptotic distribution free (ADF) method is often used to fit a proposed model. This approach uses a weight matrix based on the inverse of the matrix formed by the sample fourth moments and sample covariances. The ADF test statistic is asymptotically distributed as a chi-square variate, but its empirical performance rejects the true model too often at all but impractically large sample sizes. By comparing mean and covariance structure analysis with its peer in the multivariate linear model, we propose some modified ADF test statistics as F-tests whose distributions we approximate using F-distributions. Empirical studies show that the distributions of the new F-tests are more closely approximated by F-distributions than are the original ADF statistics when referred to chi-square distributions. Detailed analysis indicates why the ADF statistic fails on large models. An explanation for the improved behavior of Yuan and Bentler's statistic is also given. Implications for power analysis and model tests in other areas are discussed.

Key words: Mean and covariance structure, high dimensional data, F-test, Hotelling's T^2, asymptotic distribution free.

1 Introduction

High dimensional data are often collected in the social and behavioral sciences. In order to evaluate hypothesized model structures involving the relations among the observed and unobserved latent variables, as well as for dimension reduction purposes, researchers make extensive use of covariance structure analysis. Austin and Calderón (in press), Faulbaum and Bentler (1994), Hoyle (1995), and Bentler and Dudgeon (1996) provide reviews. When data obey a multivariate normal distribution, classical normal theory maximum likelihood and the corresponding likelihood ratio test will give efficient estimators and reliable inference. Since most data sets in social and behavioral sciences are not normally distributed (Micceri, 1989), researchers have to seek other methods which do not depend on the underlying distribution. The most widely known such method is the asymptotically distribution free (ADF) generalized least squares method proposed by Browne (1982, 1984) and Chamberlain (1982).

Let X_1, ..., X_n be a p-dimensional sample of size n with $\mathrm{E}X_i = \mu$ and $\mathrm{var}(X_i) = \Sigma$. In covariance structure analysis, a proposed structure $\Sigma = \Sigma(\theta)$, where θ is a q-dimensional unknown vector, is hypothesized. An important problem is to get an efficient estimator of θ and to test the adequacy of the proposed structure. Let S be the sample covariance of X_i, $vech(.)$ be an operator which transforms a symmetric matrix to a vector by picking its nonduplicate elements, $Y_i = vech[(X_i - \bar{X})(X_i - \bar{X})^T]$, and S_y be the sample covariance of Y_i. The ADF method is to model S by $\Sigma(\theta)$ with estimator $\hat{\theta}_n$ obtained by minimizing

$$F_n(\theta) = (s - \sigma(\theta))^T W_n (s - \sigma(\theta)), \tag{1.1}$$

where $W_n = S_y^{-1}$ and

$$T_n^{(1)} = nF_n(\hat{\theta}_n) \tag{1.2}$$

is its test statistic. Let $p^* = p(p+1)/2$, then under the hypothesized model the asymptotic distribution of $T_n^{(1)}$ is chi-square with $p^* - q$ degrees of freedom. Consequently, quantiles from $\chi^2_{p^*-q}$ are used to judge the significance of $T_n^{(1)}$ and correspondingly the quality of the hypothetical model. The pleasing aspect of ADF is that it gives correct inference and efficient estimators when model size is small and the sample size is large enough. However, in a factor analysis model with 15 indicators and 3 factors, the ADF method required sample size 5000 to give reliable inference in a simulation study of Hu, Bentler and Kano (1992). In a practical situation, model sizes larger than the one used by Hu et al. are not uncommon, while a sample size of more than 5000 is rarely obtainable. In small samples, the ADF method severely overrejects the true model (Chou, Bentler, & Satorra, 1991; Henly, 1993; Muthén & Kaplan, 1992), even with an unbiased weight matrix (Chan, Yung, & Bentler 1995).

More generally, the mean and covariance matrix can be modeled simultaneously. Let $Z_i = (X_i^T, vech^T(X_i X_i^T))^T$ and $\tau(\theta) = vech[\mu(\theta)\mu^T(\theta)]$. Sörbom (1974) described the case of multivariate normal data. Bentler (1989) and Muthén (1989) considered ADF estimation of models with structured means. An alternative approch was taken by Satorra (1992) and Browne and Arminger (1995), who considered modeling $\bar{Z}$ by $\xi(\theta) = (\mu^T(\theta), \sigma^T(\theta) + \tau^T(\theta))^T$. The estimator $\hat{\theta}_n$ is obtained by minimizing

$$F_n(\theta) = (\bar{Z} - \xi(\theta))^T W_n (\bar{Z} - \xi(\theta)) \tag{1.3}$$

with weight matrix $W_n = S_z^{-1}$ and S_z is the sample covariance of Z_i. The corresponding index

$$T_n^{(3)} = nF_n(\hat{\theta}_n) \tag{1.4}$$

is used as a test statistic using a chi-square distribution $\chi^2_{p+p^*-q}$ as its approximation. By looking at Z_i as response variables, mean and covariance structure analysis can be regarded as a nonlinear regression model. Yuan and Bentler (1995) proposed using the inverse of sum of cross products of the fitted residuals as a weight matrix instead of using S_z^{-1}. The corresponding statistic of Yuan and Bentler is $T_n^{(4)} = T_n^{(3)}/(1 + T_n^{(3)}/n)$, which also follows $\chi^2_{p+p^*-q}$ and is asymptotically distribution free. These authors found the empirical behavior of $T_n^{(3)}$ to be similar to that of $T_n^{(1)}$, i.e., it rejects the true model too often. They reported that $T_n^{(4)}$ gives much more reliable inference for small to intermediate sample sizes. Yuan and Bentler further proposed using the inverse of the sum of cross products of fitted residuals as a weight matrix in (1.1). The corresponding test statistic is $T_n^{(2)} = T_n^{(1)}/(1 + T_n^{(1)}/n)$, whose asymptotic distribution is also $\chi^2_{p^*-q}$. The empirical performance of $T_n^{(2)}$ was found to be similar to that of $T_n^{(4)}$, i.e., it gives more accurate inferences about model correctness in small to medium sample sizes than does the classical test $T_n^{(1)}$.

In the literature on regression, researchers often use an F-distribution to approximate a test statistic instead of a chi-square distribution when sample size is not so large. The advocates include Gallant (1975a, b) and Neill (1988) in nonlinear regression, Arnold (1980) in linear models with nonnormal errors, and many others. Since mean and covariance structure analysis is a special form of nonlinear regression, we can also use an F-distribution to approximate the distribution of the associated test statistics when sample size is not so large. Development of the relevant theory will be the main focus of this paper. By comparing covariance structure analysis with its peer in the linear model, we will also give an explanation of why $T_n^{(1)}$ and $T_n^{(3)}$ behave so badly when fitting a large model even when sample size is relatively large. A detailed analysis and explanation of the correct behavior of $T_n^{(2)}$ and $T_n^{(4)}$ will also be given.

2 Comparison of a Mean and Covariance Structure with Its Linear Peer

We shall approach the F-test of model structure via Hotelling's T^2 statistic. Let $X_1, \ldots, X_n$ be a sample from $N_n(\mu, \Sigma)$, $\bar{X}$ and S be the sample mean and sample covariance. Then Hotelling's T^2 statistic for testing $\mu = \mu_0$ is given by $T^2 = n(\bar{X} - \mu_0)^T S^{-1}(\bar{X} - \mu_0)$. More generally, let A be an $r \times p$ matrix of rank $r(< p)$, then Hotelling's statistic for testing $A\mu = b$ is given by

$$T^2 = n(A\bar{X} - b)^T(ASA^T)^{-1}(A\bar{X} - b). \tag{2.1}$$

We want to compare (2.1) with (1.4). This will be facilitated by rewriting (1.4). From Lemma 1 of Yuan and Bentler (1995), we have

$$\sqrt{n}(\bar{Z}-\xi(\hat{\theta}_n)) = \{I-\dot{\xi}(\theta_0)(\dot{\xi}^T(\theta_0)W_n\dot{\xi}(\theta_0))^{-1}\dot{\xi}^T(\theta_0)W_n\}\sqrt{n}(\bar{Z}-\xi(\theta_0))+o_p(1).$$

So we can write (1.4) as

$$T_n^{(3)} = n(\bar{Z}-\xi(\theta_0))\{W_n-W_n\dot{\xi}(\theta_0)(\dot{\xi}^T(\theta_0)W_n\dot{\xi}(\theta_0))^{-1}\dot{\xi}^T(\theta_0)W_n\}(\bar{Z}-\xi(\theta_0))+o_p(1). \tag{2.2}$$

Let $\dot{\xi}_c(\theta_0)$ be the $(p + p^*) \times (p^* - q)$ matrix whose column are orthogonal to those of $\dot{\xi}(\theta_0)$. Using Lemma 1 of Khatri (1966) and remembering that

$W_n = S_z^{-1}$, (2.2) can be further written as

$$T_n^{(3)} = n[\dot{\xi}_c^T(\theta_0)(\bar{Z} - \xi(\theta_0))]^T(\dot{\xi}_c^T(\theta_0)S_z\dot{\xi}_c(\theta_0))^{-1}[\dot{\xi}_c^T(\theta_0)(\bar{Z} - \xi(\theta_0))] + o_p(1). \tag{2.3}$$

By comparing (2.3) with (2.1), we can see that they have similar quadratic forms with (2.3) testing if there is an θ_0 such that $\dot{\xi}_c^T(\theta_0)(\xi_0 - \xi(\theta_0)) = 0$. If such a hypothesis is rejected, doubt is raised on the structure $\xi = \xi(\theta)$. In the context of covariance structure analysis, Browne (1984) gave a test statistic in a form like (2.3). Chan (1995) found that this type of statistic also rejects true models too often.

Since the scaled Hotelling statistic follows an F-distribution

$$(n-r)T^2/\{r(n-1)\} \sim F_{r,n-r}, \tag{2.4}$$

we have

$$T^2 = r(n-1)F_{r,n-r}/(n-r). \tag{2.5}$$

From (2.5), we can easily get the first and second moments of T^2. They are respectively

$$ET^2 = (n-1)r/(n-r-2), \tag{2.6}$$

$$var(T^2) = 2r(n-1)^2(n-2)/\{(n-r-2)^2(n-r-4)\}. \tag{2.7}$$

Even though the asymptotic distribution of T^2 is χ_r^2 as $n \to \infty$, the real distribution of T^2 is far from χ_r^2 for small to medium sample sizes. The critical point of T^2 is given by (3.4). To get some idea about this, consider the model in Hu et al. (1992) and Yuan and Bentler (1995). With $r = 87$ and the sample sizes used by Yuan and Bentler (1995), we list the corresponding numerical means and standard deviations of T^2 in Table 1, based on (2.6) and (2.7). The mean, standard deviation and 95% critical value of χ_{87}^2 are: 87, $\sqrt{2 \times 87} \approx 13.19$ and 109.77 respectively. Comparing with the numbers in

Table 1 and Table 5, which gives the critical values C^*_α for $\alpha = 95\%$ of T^2, we can see that if we use a chi-square distribution to approximate a Hotelling's T^2, the hypothesis will be much more highly rejected. Even though the T^2 statistic is used for testing a linear hypothesis under the assumption that the data is normal, these tables give strong evidence and an expectation that the chi-square distribution is a bad approximation to $T^{(3)}_n$ for large models with not so large sample sizes. Actually, Hotelling's T^2 test is robust to a class of distributions much larger than normal[1] (Chase & Bulgren, 1971; Mardia, 1975; Kariya, 1981). This observation motivates the use of the T^2 to approximate the distribution of $T^{(3)}_n$, or an F-distribution to approximate a rescaled $T^{(3)}_n$.

Table 1

Mean and Standard Deviation of T^2_n

Sample size	Mean	Stand deviation
150	212.51	51.03
200	155.97	31.87
300	123.28	22.32
500	105.63	17.67
1000	95.40	15.16

Similarly, the statistic $T^{(1)}_n$ can also be written in a form like (2.3). So it may also make sense to use a Hotelling's T^2 to approximate its distribution. Since the Y_i have an intercorrelation of order O(1/n), they are not totally independent. Considering further that the distribution of T^2 is robust to a data matrix whose rows are not necessarily independent (e.g. Kariya, 1981),

[1] After the current research was completed, and this manuscript was in near-final form, the second author discovered that W. Meredith (1995) also was thinking broadly about the potential relevance of Hotelling's T^2 to Browne's ADF statistic. Although his paper focused on entirely different questions from those discussed here, Meredith suggested: "Since W is sample based would it not be preferable to use a Hotelling T [sic] for evaluation? Consider the remarkable robustness of the Hotelling test". He gave no details or mathematics on this suggestion, however.

there is further motivation for using the T^2 to approximate $T_n^{(1)}$. In practice, these suggestions are most easily implemented via the F-distribution.

3 F-tests of Model Structure

In this section, we propose to scale $T_n^{(1)}$ and $T_n^{(3)}$ to create some new test statistics. Using the relation (2.4), corresponding to $T_n^{(1)}$ we have

$$T_n^{*(1)} = \{n - (p^* - q)\}T_n^{(1)}/\{(n-1)(p^* - q)\}. \tag{3.1}$$

This statistic is referred to an F-distribution with degrees of freedom $p^* - q$ and $n - (p^* - q)$. Of course, the distribution of Y_i may not fall into the class on which the Hotelling's T^2 is robust even if the data Y_i are from a family as defined in (1.3) of Kariya (1981). Since the structure $\Sigma = \Sigma(\theta)$ is nonlinear, the exact distribution of $T_n^{(1)}$ will likely not be that of a T^2, but it may be close enough in practice. Then the F-distribution should describe the distribution of (3.1). Similarly, we create a new variant of $T_n^{(3)}$ given by

$$T_n^{*(3)} = \{n - (p + p^* - q)\}T_n^{(3)}/\{(n-1)(p + p^* - q)\}. \tag{3.2}$$

Again, we use the F-distribution with degrees of freedom $p + p^* - q$ and $n - (p + p^* - q)$ to approximate the distribution of (3.2). For simplicity, we shall call $T_n^{*(1)}$ and $T_n^{*(3)}$ "F-tests". In order to see the empirical performance of $T_n^{*(1)}$ considered as an F variate, we will resort to empirical simulation. Similarly, we use empirical simulation to investigate the goodness of the approximation of $T_n^{*(3)}$ by its F variate.

The model we use here is the same as the one used by Hu et al. (1992) and Yuan and Bentler (1995), a factor analysis model $y = \Lambda f + e$ with 3 factors, each with its own 5 indicators. The number of unknown parameters in the covariance structure is 33, with $p^* = 120$, so $p * -q = 87$ for $T_n^{(1)}$. For the mean and covariance structure analysis model, we let the mean μ be a free parameter, so $q = 15 + 33 = 48$ and $p + p^* - q = 87$. Three distribution conditions were used, they are respectively: (1) both f and e are normally distributed, representing a multivariate normal distribution; (2) both f and e follow a t-distribution with 10 degrees of freedom, representing a symmetric but nonnormal distribution; (3) f are normally distributed while e follows a lognormal distribution, representing a asymmetric distribution. As in Yuan and Bentler (1995), we choose sample sizes: 150, 200, 300, 500, and 1000 respectively. For each condition, 500 replications were performed. We computed the $T_n^{*(3)}$ for modeling the mean and covariance simultaneously and the $T_n^{*(1)}$ for only covariance structure analysis. The rejection rates based on 95% quantile of the $F_{87,n-87}$ are given in Tables 2 to 4. In order to compare, we also computed the rejection rates of $T_n^{(1)}$, $T_n^{(2)}$, $T_n^{(3)}$, and $T_n^{(4)}$ based on the 95% quantile of the χ^2_{87}.

Table 2

Empirical Type I Errors For Different

Test Statistics: Normal Distribution

	Sample Size				
Statistics	150	200	300	500	1000
$T_n^{(1)}$	451/453	484/497	411	236	100
$T_n^{*(1)}$	22/453	40/497	39	42	39
$T_n^{(2)}$	0/453	11/497	20	32	35
$T_n^{(3)}$	434/436	477/496	406	234	100
$T_n^{*(3)}$	3/436	28/496	36	39	39
$T_n^{(4)}$	0/436	7/496	19	30	35

Table 3

Empirical Type I Errors For Different

Test Statistics: Multivariate t-distribution

	Sample Size				
Statistics	150	200	300	500	1000
$T_n^{(1)}$	449/450	485/498	415	232	89
$T_n^{*(1)}$	14/450	27/498	31	33	30
$T_n^{(2)}$	0/450	3/498	19	22	25
$T_n^{(3)}$	444/445	481/497	406	229	88
$T_n^{*(3)}$	2/445	21/497	27	31	28
$T_n^{(4)}$	0/445	2/497	13	22	24

Table 4

Empirical Type I Errors For Different

Test Statistics: Asymmetric Distribution

	Sample Size				
Statistics	150	200	300	500	1000
$T_n^{(1)}$	480/481	485/499	389	201	88
$T_n^{*(1)}$	10/481	15/499	20	28	34
$T_n^{(2)}$	0/481	1/499	7	14	26
$T_n^{(3)}$	458/460	480/495	383	200	91
$T_n^{*(3)}$	1/460	10/495	16	26	33
$T_n^{(4)}$	0/460	0/495	6	15	26

From Tables 2 to 4, we can see that the empirical type I errors of $T_n^{(1)}$ and $T_n^{(3)}$ are so large that they can not be used in practice. The empirical type I errors of our F-tests $T_n^{*(1)}$ and $T_n^{*(3)}$ are a little over the nominal errors in most of the cases for normal and symmetric data. For skewed data in Table 4, the rejection rates of these F-tests are less than the nominal rates for sample sizes 200 (counting the unconverged samples as rejection) and 300. Based on the 500 replications, the statistics $T_n^{(2)}$ and $T_n^{(4)}$ give the smallest rejection rates in all the cases studied. Overall, the performances of $T_n^{*(1)}$,

$T_n^{*(3)}$, $T_n^{(2)}$ and $T_n^{(4)}$ are much better than those of $T_n^{(1)}$ and $T_n^{(3)}$ with $T_n^{(2)}$ and $T_n^{(4)}$ being nearer the nominal rate. Note that $T_n^{(1)} > T_n^{(2)}$ and $T_n^{(3)} > T_n^{(4)}$ numerically.

The significance of $T_n^{(2)}$, $T_n^{(4)}$ and $T_n^{*(1)}$, $T_n^{*(3)}$ depends on the critical values from chi-square and F-distributions respectively, so we need numerical methods to compare these statistics. Since $T_n^{(2)}$ and $T_n^{(4)}$ use chi-square distributions as their approximations, the critical values are given by

$$C_\alpha = n\chi_r^2(\alpha)/\{n - \chi_r^2(\alpha)\}, \tag{3.3}$$

with $r = p^* - q$ and $r = p + p^* - q$ corresponding to $T_n^{(2)}$ and $T_n^{(4)}$ respectively, where $\chi_r^2(\alpha)$ is the upper α critical value of χ_r^2. Since $T_n^{*(1)}$ and $T_n^{*(3)}$ use F-distributions as their approximations, from (3.1) and (3.2), their critical values are given by

$$C_\alpha^* = r(n-1)F_{r,n-r}(\alpha)/(n-r), \tag{3.4}$$

with $r = p^* - q$ and $r = p + p^* - q$ respectively. For $r = 87$, which is used in the empirical studies in Hu et al (1992) and Yuan and Bentler (1995), we list some of the C_α and C_α^* for $\alpha = 95\%$ in Table 5 based on (3.3) and (3.4) for selected sample sizes. Hu et al. reported that the behavior of the ADF statistic tends to nominal when sample sizes are around 5000. Comparing the results from Table 5 with the 95% critical value of χ_{87}^2, which is 109.77, we can see why the sample size requirement for $T_n^{(1)}$ is so large. So when sample sizes are around 5000, the three types of test statistics will give approximately the same rejection rates for the factor analysis model with r=87. When sample sizes are less than 500, there will be some differences between the test statistics $T_n^{*(1)}$, $T_n^{*(3)}$ and $T_n^{(2)}$, $T_n^{(4)}$, with the rejection rates of the F-tests $T_n^{*(1)}$ and $T_n^{*(3)}$ necessarily being a little bit higher.

Table 5

Critical Values C_α and C^*_α For $\alpha = 95\%$

	Sample Size						
	150	200	300	500	1000	3000	5000
C_α	409.33	243.33	173.12	140.65	123.31	113.94	112.24
C^*_α	305.23	212.95	162.65	136.51	121.72	113.49	111.98

So far our attention has been focused on the tail probability, which is important for testing purposes. Sometimes, we may have interest in the whole distribution of a statistic. For this, we rely on the Kolmogorov-Smirnov (K-S) statistic to see the quality of the approximations to these different test statistics. Let $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ be an ordered sample from a continuous distribution. The empirical distribution function $F_n(x)$ is defined by

$$F_n(x) = \begin{cases} 0, & x < X_{(1)} \\ i/n, & X_{(i)} \leq x < X_{(i+1)} \\ 1, & X_{(n)} \leq x \end{cases}$$

Suppose we want to test if a sample is from a population whose distribution function is $F(x)$. The K-S test statistic is given by

$$D_{KS} = \sup_x |F_n(x) - F(x)|.$$

Easy to approach references on the K-S statistic can be found in Birnbaum (1952), Gibbons (1985, sections 4.4-4.6), and Stuart and Ord (1991, §30.37-§30.42). The 95% and 99% critical values of D_{KS} based on its asymptotic distribution are $1.3581/\sqrt{n}$ and $1.6276/\sqrt{n}$. For n=500, these critical values are approximately 0.0607 and 0.0728.

Based on our empirical studies described earlier, the D_{KS} was calculated for each case. The statistics are listed in Tables 6 to 8, corresponding to Tables 2 to 4. From these numbers, we can see that for sample size 150 the K-S statistics corresponding to $T_n^{(1)}$ and $T_n^{(3)}$ are almost 10 times larger than those of $T_n^{*(1)}$ and $T_n^{*(3)}$, and about 6 times larger than those of $T_n^{(2)}$ and $T_n^{(4)}$. The K-S statistics of our F-tests $T_n^{*(1)}$ and $T_n^{*(3)}$ are the smallest for all sample sizes presented here. For sample sizes 300, 500, and 1000, the K-S statistics corresponding to $T_n^{(2)}$ and $T_n^{(4)}$ are almost as good as those of $T_n^{*(1)}$ and $T_n^{*(3)}$ for normal data and symmetric data; for the skewed data, those of $T_n^{(2)}$, $T_n^{(4)}$ and $T_n^{*(1)}$, $T_n^{*(3)}$ are similar when sample sizes are 500 and 1000. Five of the K-S statistics corresponding to $T_n^{*(1)}$ are not significant under the 99% critical value, five for $T_n^{*(3)}$, one for $T_n^{(2)}$, and one for $T_n^{(4)}$. Note that since there exist some nonconverging samples for sample sizes 150 and 200, the statistics corresponding to the converged samples may not be independent. So for sample sizes 150 and 200, the statistics are given only for exploratory purposes and to provide reference values.

Table 6

K-S Statistics For Different

Test Statistics: Normal Distribution

	Sample Size				
Statistics	150	200	300	500	1000
$T_n^{(1)}$	.9886	.9384	.7926	.5375	.2845
$T_n^{*(1)}$	.0746	.1269	.1371	.1072	.0534
$T_n^{(2)}$	.1640	.1689	.1577	.1107	.0600
$T_n^{(3)}$	.9853	.9343	.7890	.5314	.2833
$T_n^{*(3)}$	.0914	.0922	.1275	.1014	.0528
$T_n^{(4)}$	.1796	.1459	.1498	.1060	.0592

Table 7

K-S Statistics For Different

Test Statistics: Multivariate t-distribution

	Sample Size				
Statistics	150	200	300	500	1000
$T_n^{(1)}$	.9892	.9326	.7869	.5718	.3022
$T_n^{*(1)}$	.0619	.1144	.1292	.1243	.0723
$T_n^{(2)}$	.1557	.1461	.1557	.1376	.0778
$T_n^{(3)}$	.9871	.9286	.7807	.5691	.3012
$T_n^{*(3)}$	.1263	.0873	.1172	.1193	.0708
$T_n^{(4)}$	.2009	.1309	.1422	.1334	.0760

Table 8

K-S Statistics For Different

Test Statistics: Asymmetric Distribution

	Sample Size				
Statistics	150	200	300	500	1000
$T_n^{(1)}$	.9897	.9314	.7606	.5191	.3200
$T_n^{*(1)}$	.0695	.0553	.0782	.0734	.0948
$T_n^{(2)}$	.1683	.1198	.1141	.0840	.1004
$T_n^{(3)}$	.9905	.9249	.7578	.5254	.3195
$T_n^{*(3)}$	.1448	.0701	.0685	.0723	.0949
$T_n^{(4)}$	.2183	.1321	.1050	.0891	.1028

Comparing the results with those in Tables 2 to 4, we can see that there exists some discord between the measures of tail probability and the K-S measures of distributional misfit. Approximations of our F-tests $T_n^{*(1)}$ and $T_n^{*(3)}$ by F-distributions are always the best judging by the K-S statistic, while they are not as good as $T_n^{(2)}$ and $T_n^{(4)}$ in overall approximations of tail probabilities. Considering that most of the time these statistics are used for testing purposes only, their tail probabilities are more important from a practical point of view. An approximation to a distribution may not be universally good everywhere, may be good at the middle and bad at the tails or vice versa. For example, the direct Edgeworth expansion usually gives a very good approximation at the center of a distribution but can be very bad regarding tail probabilities (Barndorff-Nielsen & Cox, 1989, Chapter 4; Field & Ronchetti, 1990), while the saddle point approximation focuses on improving the approximation around a point of interest. Since the K-S statistic measures the distance between $F_n(x)$ and $F(x)$, and the tail probability compares a sample with a specific quantile, it is not surprising that the differences exist.

4 An Explanation for Yuan and Bentler's Statistic

From the empirical evidence in last section, the Hotelling's T^2 distribution, as transformed into an F variate, gives a much better approximation to the behavior of $T_n^{(1)}$ and $T_n^{(3)}$ than the large sample size based chi-square distribution. The statistics $T_n^{(2)}$ and $T_n^{(4)}$ perform equally well when using chi-square distributions as their approximations. We give an explanation for these divergent results based on the T^2 distribution.

Let X_r^2 and X_{n-r}^2 be independent and follow chi-square distributions with degrees of freedom r and $n-r$ respectively. Then

$$F_{r,n-r} = \frac{X_r^2/r}{X_{n-r}^2/(n-r)}$$

follows an F-distribution with degrees of freedom r and $n-r$. It follows from (2.5)

$$T^2 = \frac{(n-1)X_r^2}{X_{n-r}^2}. \tag{4.1}$$

So

$$\frac{T^2}{1+T^2/n} = \frac{X_r^2}{X_n^2/n - X_{n-r}^2/\{n(n-1)\}}. \tag{4.2}$$

Rewrite (4.1) as

$$T^2 = \frac{X_r^2}{\frac{(n-r)}{(n-1)}\frac{X_{n-r}^2}{(n-r)}} \tag{4.3}$$

for easy comparison. Since the numerators in (4.2) and (4.3), the χ_r^2 variates used as their approximations are the same, the qualities of the approximations are really decided by the denominators. If a denominator equals 1, then the approximation is perfect. As $X_m^2/m \rightarrow 1$ when m increases, by comparing (4.2) and (4.3), we can see that the denominator in (4.2) not only recovers the degrees of freedom r but also changes the bias from a multiplicative factor $(n-r)/(n-1)$ to a minus factor of $X_{n-r}^2/n(n-1)$. Even when a sample size n is very large, if r is not so small as in most practical models, the bias brought in by $(n-r)/(n-1)$ can be overwhelming. For a fixed n, the amount of bias in (4.3) will increase as r increases. On the other hand, for each fixed n, the bias brought in by $X_{n-r}^2/n(n-1)$ will decrease as model sizes increases. The maximum amount of bias is approximately $1/n$ when $r = 1$.

Even though our explanation of the properties of $T_n^{(2)}$ and $T_n^{(4)}$ is based on the T^2 distributions, these properties are reflected in Tables 2 to 4 and 6 to 8. Furthermore, the empirical means and standard deviation of $T_n^{(1)}$ and $T_n^{(3)}$ as reported in Hu et al (1992) and Yuan and Bentler (1995) are also near those corresponding to Hotelling's T^2, as shown in Table 1.

5 Discussion

Outside the standard linear model, the distributions of most goodness of fit test statistics are approximated by chi-square distributions. These approximations are supported by large sample theory. However, these chi-square approximations can be very bad, especially when the models are very large. This problem occurs even when sample sizes are fairly large. In covariance structure analysis, which is usually used for high dimensional data analysis, this problem becomes obviously serious. Even though most researchers are aware of this problem, they still continue to use a chi-square approximation because of the lack of more reliable alternatives. Yung and Bentler (1994) used a bootstrap method to improve the performance of $T_n^{(1)}$ which is computationly intensive. Yuan and Bentler (1995) proposed statistics $T_n^{(2)}$ and $T_n^{(4)}$ which do not need extra computation beyond that of $T_n^{(1)}$ and $T_n^{(3)}$. This paper furthermore proposes the F-tests $T_n^{*(1)}$ and $T_n^{*(3)}$ and proposes the use of F-distributions to approximate their distributions. Our approach is motivated by the resemblance of $T_n^{(1)}$ and $T_n^{(3)}$ to the Hotelling T^2. Empirical evidence shows that the F-distribution approximations are much better than the large sample theory based chi-square approximations to the distributions of $T_n^{(1)}$ and $T_n^{(3)}$. The K-S statistics also suggest the reasonableness of the F-distribution approximations. As compared with $T_n^{(2)}$ and $T_n^{(4)}$, the rejection rates of $T_n^{*(1)}$ and $T_n^{*(3)}$ are a little bit higher for sample sizes 300 to 1000 for symmetric data. For the asymmetric data, on the other hand, the rejection rates of $T_n^{*(1)}$ and $T_n^{*(3)}$ perform better than those of $T_n^{(2)}$ and $T_n^{(4)}$ for sample sizes under 500.

In this paper we have only investigated the test statistics and their distributions when hypothetical structures are correct. Under alternative hypotheses, noncentral chi-squares have been used to describe the distributions of $T_n^{(1)}$ and $T_n^{(3)}$. From our limited experience in empirical study, the powers of $T_n^{(1)}$ and $T_n^{(3)}$ are almost always 1 under a small departure from the null hypothesis for any sample size. This is because for large sample sizes, the noncentrality parameters are very large and the power should be approximately 1; while for small and medium sample sizes, the powers are almost equal to 1 even if the null hypotheses are correct! A reason behind this is that noncentral chi-square variates are bad approximations to the distributions of $T_n^{(1)}$ and $T_n^{(3)}$ under alternative hypotheses. Since Hotelling's T^2 gives reasonable approximations to the distributions of $T_n^{(1)}$ and $T_n^{(3)}$, under alternative hypotheses we also may approximate their distributions by noncentral Hotelling's T^2's and, consequently, approximate the distributions of $T_n^{*(1)}$ and $T_n^{*(3)}$ by noncentral F-distributions. Such approximations will have asymptotically the same power as noncentral chi-square approximations to those of $T_n^{(1)}$ and $T_n^{(3)}$, but there will be some differences for small to medium sample sizes. This will have a measurable effect on the uses to which noncentral distrubutions are put, for example, power analysis (e.g., Saris & Satorra, 1993) and practical measures of fit such as the comparative fit index (e.g., Bentler, 1989) (see also Hoyle, 1995). More informative conclusions need further investigation and will be given elsewhere.

In this paper we have addressed only the goodness-of-fit χ^2 test for evaluating model structure. In practice, however, researchers also evaluate sets of restrictions via χ^2 difference tests, Lagrange Multiplier tests, and Wald tests (e.g., Bentler & Dijkstra, 1985; Satorra, 1989). The statistics involved in these approaches are treated as asymptotic chi-square variates, and for the reasons enumerated above we propose that our approach based on F-tests may provide a more accurate evaluation of hypothesized restrictions. This topic will be addressed in a separate paper.

We have concentrated our development on tests associated with efficient estimators. They also are relevant to nonefficient estimators. Consider, for example, Browne's (1984) test for a nonefficient estimator, e.g. the least squares estimator. As noted previously, this is of the form $T_n^{(3)}$, and Chan (1995) found that it rejected true models too frequently. Clearly, our F-test variant of $T_n^{(3)}$, that is $T_n^{*(3)}$, should also be a better test of model structure at most realistic sample sizes than Browne's original statistic for nonefficient estimators. This suggestion will be evaluated fully elsewhere.

Even though our simulation studies are limited to mean and covariance structure analysis, these types of test statistics can also be applied to other areas in which the chi-square approximations reject the true models too often. An obvious extention is to multiple-sample ADF theory (e.g., Bentler, Lee, and Weng, 1987; Muthén, 1989). Our statistics also should apply to categorical variables methods, since it has been reported that Muthén's (1987) LISCOMP and Jöreskog and Sörbom's (1993) LISREL test statistics over-reject true models (e.g., Bentler, 1994; Dolan, 1994). Similarly, there are usually a large number of parameters in principal component analysis, in panel data analysis, and in log-linear models. If in these areas the rejection rates of some χ^2 approximations are higher than they should be, our proposed test statistics may perform better.

References

[1] Arnold, S. F. (1980). Asymptotic validity of F-tests for the ordinary lin-

ear model and the multiple correlation model. *Journal of the American Statistical Association,* **75**, 890–849.

[2] Austin, J. T., & Calderón, R. F. (in press). Theoretical and technical contributions to covariance structure modeling: An updated annotated bibliography. *Structural Equation Modeling.*

[3] Barndorff-Nielsen, O. E., & Cox, D. R. (1989). *Asymptotic techniques for use in Statistics.* London: Chapman and Hall.

[4] Bentler, P. M. (1989). *EQS structural equations program manual.* Los Angeles: BMDP Statistical Software.

[5] Bentler, P. M. (1994). On the quality of test statistics in covariance structure analysis: Caveat emptor. In C. R. Reynolds (ed.), *Cognitive assesment: A multidisciplinary perspective* (pp. 237–260). New York: Plenum.

[6] Bentler, P. M., & Dijkstra, T. (1985). Efficient estimation via linearization in structural models. In P. R. Krishnaiah (ed.), *Multivariate analysis VI* (pp. 9–42). Amsterdam: North-Holland.

[7] Bentler, P. M., & Dudgeon, P. (1996). Covariance structure analysis: Statistical practice, theory, and directions. *Annual Review of Psychology,* **47**, 563–592.

[8] Bentler, P. M., Lee, S.-Y., & Weng, J. (1987). Multiple population covariance structure analysis under arbitrary distribution theory. *Communications in Statistics -Theory,* **16**, 1951–1964.

[9] Birnbaum, Z. W. (1952). Numerical tabulation of the distribution of Kolmogorov's statistic for finite sample size. *Journal of the American Statistical Association,* **47**, 425–441.

[10] Browne, M. W. (1982). Covariance structure analysis. In D. M. Hawkins (ed.), *Topics in applied multivariate analysis* (pp. 72–141). England: Cambridge University Press.

[11] Browne, M. W. (1984). Asymptotic distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology,* **37**, 62–83.

[12] Browne, M. W., & Arminger G. (1995). Specification and estimation of mean and covariance models. In G. Arminger, C. C. Clogg, & M. E. Sobel (eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 185–249). New York: Plenum.

[13] Chamberlain, G. (1982). Multivariate regression models for panel data. *Journal of Econometrics,* **18**, 5–46.

[14] Chan, W. (1995). *Covariance structure analysis of ipsative data.* Ph.D. thesis, University of California, Los Angeles.

[15] Chan, W., Yung, Y.-F., & Bentler, P. M. (1995). A note on using an unbiased weight matrix in the ADF test statistic. *Multivariate Behavioral Research,* **30**, 453–459.

[16] Chase, G. R., & Bulgren, W. G. (1971). A Monte Carlo investigation of the robustness of T^2. *Journal of the American Statistical Association,* **66**, 499–502.

[17] Chou, C.-P., Bentler, P. M., & Satorra, A. (1991). Scaled test statistics and robust standard errors for nonnormal data in covariance structure analysis: A Monte Carlo study. *British Journal of Mathematical and Statistical Psychology,* **44**, 347–357.

[18] Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology,* **47**, 309–326.

[19] Faulbaum, F., & Bentler, P. M. (1994). Causal modeling: Some trends and perspectives. In I. Borg & P. P. Mohler (eds.), *Trends and perspectives in empirical social research* (pp. 224–249). Berlin: Walter de gruyter.

[20] Field, C., & Ronchetti, E. (1990). *Small sample asymptotics.* Lecture notes-monograph series, 13, Hayward, CA: Institute of Mathematical Statistics.

[21] Gallant, A. R. (1975a). The power of the likelihood ratio test of location in nonlinear regression models. *Journal of the American Statistical Association,* **70**, 198–203.

[22] Gallant, A. R. (1975b). Testing a subset of the parameters of a nonlinear regression model. *Journal of the American Statistical Association,* **70**, 927–932.

[23] Gibbons, J. D. (1985). *Nonparametric statistical inference, second edition.* New York: Marcel Dekker, Inc..

[24] Henly, S. J. (1993). Robustness of some estimators for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology,* **46**, 313–338.

[25] Hoyle, R. (ed) (1995). *Structural equation modeling: Concepts, issues, and applications.* Thousand Oaks, CA: Sage.

[26] Hu, L., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin,* **112**, 351–362.

[27] Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8 user's reference guide.* Chicago: Scientific Software International.

[28] Kariya, T. (1981). A robustness property of Hotelling's T^2-test. *The Annals of Statistics,* **9**, 211–214.

[29] Khatri, C. G. (1966). A note on a MANOVA model applied to problems in growth curve. *Annals of the Institute of Statistical Mathematics,* **18**, 75–86.

[30] Mardia, K. V. (1975). Assessment of multinormality and the robustness of Hotelling's T^2 test. *Applied Statistics,* **24**, 163–171.

[31] Meredith, W. (1995, October). Alternative fit functions. Paper presented at Society of Multivariate Experimental Psychology, Blaine, WA.

[32] Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin,* **105**, 156–166.

[33] Muthén, B. (1987). *LISCOMP: Analysis of linear structural equations using a comprehensive measurement model.* Mooresville, IN: Scientific Software.

[34] Muthén, B. (1989). Multiple group structural modelling with non-normal continuous variables. *British Journal of Mathematical and Statistical Psychology,* **42**, 55–62.

[35] Muthén, B., & Kaplan, D. (1992). A comparison of some methodologies for the factor analysis of non-normal likert variables: A note on

the size of the model. *British Journal of Mathematical and Statistical Psychology,* **45**, 19–30.

[36] Neill, J. W. (1988). Testing for lack of fit in nonlinear regression. *The Annals of Statistics,* **16**, 733–740.

[37] Saris, W., & Satorra, A. (1993). Power evaluations in structural equation models. In K. A. Bollen & J. S. Long (eds.), *Testing structural equation models* (pp. 181–204). Newbury Park, CA: Sage.

[38] Satorra, A. (1989). Alternative test criteria in covariance structure analysis: A unified approach. *Psychometrika,* **54**, 131–151.

[39] Satorra, A. (1992). Asymptotic robust inferences in the analysis of mean and covariance structures. In P. V. Marsden (ed.), *Sociological methodology* (pp. 249–278). Blackwell: Oxford.

[40] Sörbom, D. (1974). A general method for studying differences in factor means and factor structures between groups. *British Journal of Mathematical and Statistical Psychology,* **27**, 229–239.

[41] Stuart, A., & Ord, J. K. (1991). *Kendall's Advanced Theory of Statistics, Vol. 2, 5th ed.*. New York: Oxford University Press.

[42] Yuan, K.-H., & Bentler, P. M. (1995). Mean and covariance structure analysis: Theoretical and practical improvement. Under editorial review. UCLA Statistics Series No. 194, Center for Statistics.

[43] Yung, Y. F., & Bentler, P. M. (1994). Bootstrap-corrected ADF test statistics in covariance structure analysis. *British Journal of Mathematical and Statistical Psychology,* **47**, 63–84.